ISSN 1392-0561. INFORMACIJOS MOKSLAI. 2007 42–43

# Asmenvardžių ir vietovardžių rašybos klaidos pokalbių programų registracijos duomenyse

**Gintautas Grigas**

Matematikos ir informatikos instituto docentas, daktaras
Institute of Mathematics and Informatics, Assoc. Prof., PhD
Akademijos g. 4, LT-08663 Vilnius
Tel. (+370 5) 210 93 44
El. paštas: grigas@ktl.mii.lt

**Sigita Pedzevičienė**

Matematikos ir informatikos instituto inžinierė programuotoja
Institute of Mathematics and Informatics, Engineer programmer
Akademijos g. 4, LT-08663 Vilnius
Tel. (+370 5) 210 93 44
El. paštas: sigitab@ktl.mii.lt

*Pateikiama vardų, pavardžių ir miestų rašybos klaidų statistika pokalbių sistemos „Skype“ abonentų registracijos duomenyse. Tikrinamos dažniausiai pasitaikančios klaidos – raidžių, nesančių pagrindinėje lotyniškos abėcėlės dalyje, pakeitimas kitomis, į jas panašiomis raidėmis be diakritinių ženklų. Palyginama, kaip šias klaidas daro lietuviai, čekai, danai, estai, islandai, latviai, lenkai, vokiečiai. Tikrinimui buvo imama po šešis kiekvienos kalbos dažniau vartojamus vardus ir po tris kiekvienos valstybės didžiausių miestų pavadinimus, turinčius raidžių su diakritiniais ženklais. Nustatyta, kad praktiškai klaidų nedaro danai ir vokiečiai, o daugiausia daro lietuviai. Bandoma ieškoti klaidų priežasčių keliant hipotezes, dėl ko tokios klaidos daromos.*

Vis daugiau kalbama apie kompiuterinį raštingumą. Koks jo santykis su paprastu raštingumu? Ar pavyksta vieną derinti su kitu? Kokia mūsų bendravimo kultūra virtualioje erdvėje? Kaip mes prisistatome kitiems? Ar tinkamai prisistatome? Kaip užrašome savo gyvenvietės (miesto) pavadinimą? Preliminarūs eksperimentai su asmenvardžiais (Grigas, 2007) parodė, kad klaidų daroma nemažai, ir tai buvo paskata atlikti čia pateikiamą išsamesnę analizę.

Buvo tikrinama raidžių, nesančių pagrindinėje lotyniškos abėcėlės dalyje (faktiškai anglų kalbos abėcėlėje) pakeitimas kitomis, į jas panašiomis raidėmis be diakritinių ženklų. Palyginama, kaip šias klaidas daro lietuviai ir mūsų kaimynai, vartojantys lotynišką abėcėlę: latviai, estai, lenkai, čekai, danai, vokiečiai, taip pat islandai, kurių pavardžių rašymas turi palyginimui patogią bendrą savybę – abiejų tautų visos taisyklingai sudarytos moterų pavardės turi raidžių su diakritiniais ženklais (lietuvių -ytė, -aitė, -iūtė, islandžių -dóttir).

Analizei pasirinkome registravimosi duomenis pokalbių programoje (sistemoje) „Skype“ dėl to, kad šia sistema naudojasi įvairūs visuomenės sluoksniai, sistema turi gana daug abonentų – apie 196 milijonus (Courtney, 2007), programa populiari (išversta į 27 kalbas, tarp jų – lietuvių, estų, lenkų, čekų, danų, vokiečių), registracijos duomenys koduojami Unikodu, juos galima nesunkiai rasti visų (arba pakankamai didelio skaičiaus) prisijungusiųjų prie sistemos.

Sistemos „Skype“ duomenų bankas į užklausą duoda ribotą skaičių radinių (abonentų) – apie 90. Todėl užklausas skaidėme pagal amžiaus grupes (jų yra šešios) ir lytį. Šitaip išskaidę užklausą į 12 dalių galėjome gauti maždaug iki tūkstančio radinių. Daugeliu atvejų tiek radinių faktiškai nebuvo, todėl galima tvirtinti, kad į apskaitą patekdavo visi analizuojamą savybę

*1 lentelė.* ***Vardų rašyba***

| Kalba | Vardai | Iš viso | Taisyklingų | Taisyklingų % |
|---|---|---|---|---|
| Čekų | Dáša, Dušan, Luboš, Pavlína, Václav, Věra | 665 | 486 | 73 |
| Danų | Søren, Jørgen, Bjørg, René, Lækker, Børge | 379 | 313 | 83 |
| Estų | Väike, Külliki, Ülle, Jüri, Märten, Tõnu | 334 | 284 | 85 |
| Islandų | Ævar, Davíð, Ólöf, Rósa, Sigurður, Úlfar | 303 | 197 | 65 |
| Latvių | Ināra, Jānis, Ainārs, Ervīns, Irēne, Līga | 407 | 121 | 30 |
| Lenkų | Michał, Paweł, Rafał, Aśka, Małgorzata, Łukasz | 700 | 279 | 40 |
| Lietuvių | Aušra, Ramunė, Rūta, Šarūnas, Žilvinas, Kęstutis | 634 | 98 | 15 |
| Vokiečių | Bärbel, Dörte, Jürgen, Rüdiger, Jörg, Björn | 671 | 646 | 96 |

turintys abonentai. Tik atskirais atvejais analizuojant didelių miestų rašybą abonentų skaičius šią ribą viršijo, pavyzdžiui, latvių, gyvenančių Rygoje ir taisyklingai (arba netaisyklingai) rašančių žodį „Rīga“ (t. y. ne „Riga“). Tačiau tai neturėjo didesnės įtakos tikslumui, nes programa „Skype“ pateikiamus abonentus rikiuoja pagal asmenvardžius, neskirdama miestų pavadinimuose raidžių su diakritiniais ženklais nuo į jas panašių raidžių be diakritinių ženklų (šiuo atveju „ī“ nuo „i“). Taigi, ribotas radinių skaičius ribojo tik imtį, kuri esti ir taip didesnė už kitas imtis, į kurias patenka visi abonentai.

Atkreipsime dėmesį, kad analogiškus ribojimus turi ir kitos pokalbių programos.

Analizei ėmėme duomenis tik tų asmenų, kurie gyvena jų deklaruotoje valstybėje ir kalba tos valstybės kalba, gyvena tame mieste, kurio rašybą nagrinėjame.

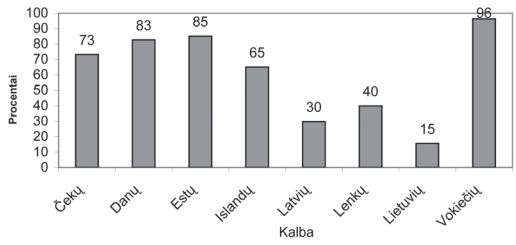
Skaičiavome aktyvius (prisijungusius) abonentus 2007 metų balandį per darbo dienas.

## Vardų rašybos klaidos

Registruojantis programa prašo pateikti asmenvardį, t. y. vardą ir pavardę. Nustatyti, ar gramatiškai taisyklingai užrašytos net ir lietuviškos pavardės, ne visada galima. Pavyzdžiui, jeigu matome Siskus, tai dar neaišku, ar iš tikrųjų Siskus, ar „šveplai“ užrašytas Šiskus, Siškus ar Šiškus. Dar sudėtingiau su užsieniečių pavardėmis. Todėl išsamiai analizavome tik vardus. Parinkome po šešis kiekvienos kalbos vardus, turinčius bent vieną raidę su diakritiniu ženklu ir dažniau naudojamus sistemoje „Skype“.

Analizės rezultatai pateikiami 1 lentelėje, o grafiškai pavaizduoti 1 paveiksle.

Matome, kad raštingiausi yra vokiečiai ir tik 15% lietuvių savo vardus užrašo taisyklingai.




*1 pav.* ***Taisyklingai parašytų vardų palyginimo diagrama***

## Pavardžių rašyba

Lietuviškos moterų pavardės baigiasi raide ė (-ytė, -aitė, -iūtė, -ienė), taigi kiekviena pavardė turi raidę ė.

Pavardžių rašymo dėsningumą turi ir islandų kalba. Visos islandžių pavardės baigiasi „-dóttir“ (kas reiškia „duktė“) ir turi vieną raidę, nesančią anglų kalbos abėcėlėje, – ó (o su dešininiu kirčiu). Pavyzdžiui, Gunnars dukterų pavardės būtų Gunnarsdóttir (o sūnų – Gunnarsson).

Pasinaudojant šiuo panašumu, galima lengvai palyginti, kaip pasirašo lietuvės ir islandės. Suskaičiavę gavome, kad taisyklingai pavardes pasirašo apie 20% lietuvių ir apie 80% islandžių.

Kokią išvadą galima daryti iš to, kad abiejų šalių moterys pavardes parašo truputį geriau negu vardus? Manome, kad tai yra dėl to, kad

buvo skaičiuojamos tik tos, kurios pavardes nurodė. O jos tvarkingesnės už tas, kurios nori likti anoniminės.

## Miestų pavadinimų rašyba

Iš kiekvienos valstybės parinkome po tris didesnius miestus, kurių pavadinimuose yra raidžių su diakritiniais ženklais. Gavome rezultatus, pateikiamus 2 lentelėje ir grafiškai vaizduojamus 2 paveiksle. Miestų pavadinimai pateikiami originalo kalba.

Klaidų beveik nedaro danai ir vokiečiai, labai mažai – estai. Daugiausia klaidų daro klaipėdiškiai, o antroje vietoje nuo galo būtų rygiečiai.

Diagramoje kiekvienos valstybės miestus išdėstėme jų mažėjimo (pagal visų prisijungusiųjų skaičių) eile. Iš diagramos matyti, kad beveik visų mažesnių miestų gyventojai daro mažiau klaidų negu tos pačios valstybės didesnių miestų gyventojai. Ne taip išsirikiavo tik Krokuva ir Gdanskas, matyt, dėl to, kad ir jų dydžiai panašūs.

*2 lentelė.* ***Miestų pavadinimų rašyba***

| Valstybė | Miestas | Iš viso | Taisyklingų | Teisyklingų % | Bendras % |
|---|---|---|---|---|---|
| Čekija | Plzeň | 494 | 423 | 86 | 87 |
| | Děčin | 155 | 140 | 90 | |
| | Tábor | 139 | 125 | 90 | |
| Danija* | København | 512 | 512 | 100 | 100 |
| | Århus | 198 | 198 | 100 | |
| | Sønderborg | 37 | 37 | 100 | |
| Estija | Pärnu | 146 | 140 | 96 | 97 |
| | Túri | 17 | 17 | 100 | |
| | Pölva | 15 | 15 | 100 | |
| Islandija | Kópavagur | 117 | 73 | 62 | 68 |
| | Hafnarfjörður | 80 | 51 | 64 | |
| | Garðabær | 31 | 31 | 100 | |
| Latvija | Rīga | 862 | 137 | 16 | 23 |
| | Liepāja | 304 | 103 | 34 | |
| | Cēsis | 134 | 54 | 40 | |
| Lenkija** | Łódź | 555 | 448 | 81 | 82 |
| | Kraków | 514 | 430 | 84 | |
| | Gdańsk | 416 | 335 | 81 | |
| Lietuva | Klaipėda | 649 | 75 | 12 | 17 |
| | Šiauliai | 420 | 92 | 22 | |
| | Panevėžys | 303 | 66 | 22 | |
| Vokietija*** | München | 682 | 679 | 100 | 100 |
| | Köln | 397 | 395 | 99 | |
| | Nürnberg | 179 | 179 | 100 | |

* Raidė Å ir dviraidis AA danų kalboje yra lygiaverčiai. Dviraidį rašė apie 20% Århus miesto gyventojų. Jų į analizės rezultatus neįtraukėme (faktiškai jie ir nepakeistų raštingumo procento).

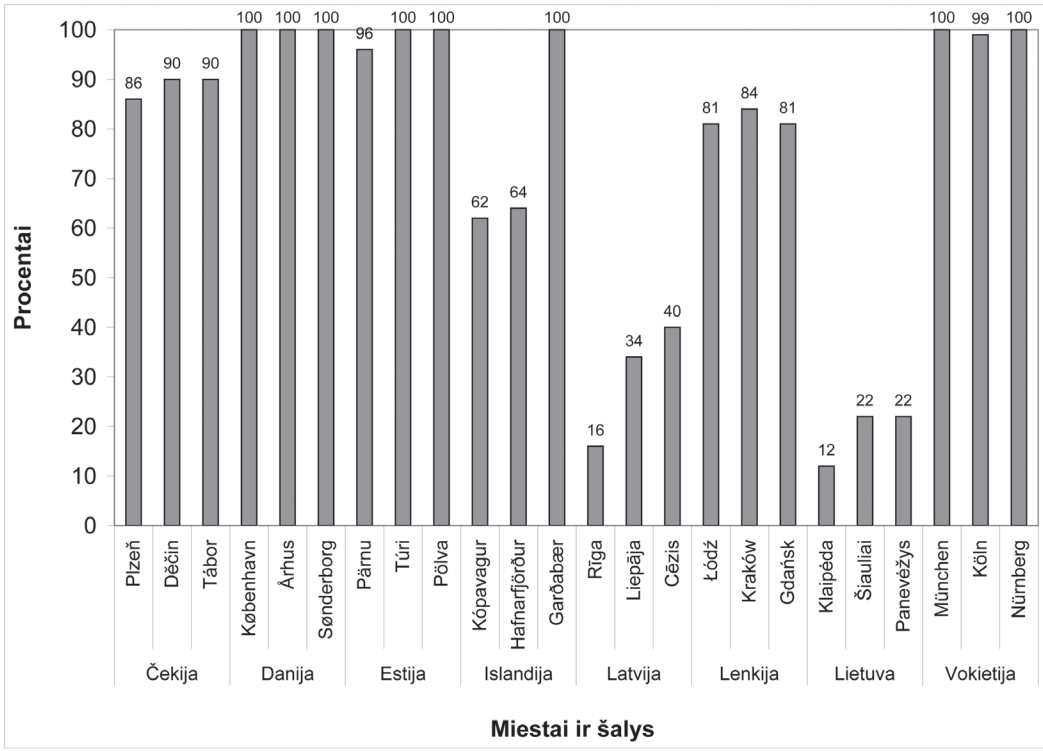
** Sistemos „Skype“ paieška raidžių L ir Ł netapatina. Todėl abiem raidėms buvo daromos atskiros paieškos.

*** Vokiečių kalbos gramatika leidžia raides ä, ö, ü ir ß pakeisti raidžių poromis ae, oe, ue ir ss. Šia taisykle pasinaudojo apie 8% vokiečių, užrašydami trijų nagrinėtų miestų pavadinimus. Jų į analizės rezultatus neįtraukėme.

## Išvados

1. Lietuviai daro gerokai daugiau asmenvardžių ir miestų pavadinimų rašybos klaidų negu bet kuri nagrinėta tauta (čekai, danai, estai, islandai, latviai, lenkai, vokiečiai), vartojanti lotyniškus rašmenis.

2. Žemo lietuvių raštingumo lygio priežastis turėtų būti susijusi su rašto ženklų vartojimu kompiuteryje: jų kodavimu, rinkimu. Visų nagrinėtų kalbų koduotės šiuolaikinėse operacinėse sistemose („Windows“, „Linux“) yra sudarytos pagal tą patį principą ir atitinka tarptautinius standartus. Todėl kodavimas neturėtų daryti įtakos klaidų kiekiui. Kitose šalyse naudojamos klaviatūros yra sudarytos pagal tą patį principą, nustatytą tarptautinio standarto (ISO, 1994). Lietuvoje naudojama „skaičiukinė“ klaviatūra standartų neatitinka, ja sunkiau surinkti raides su diakritiniais ženklais. Todėl ji laikytina klaidų priežastimi.




*2 p a v.* ***Taisyklingai parašytų miestų pavadinimų palyginimo diagrama***

## LITERATŪRA


- COURTNEY, J. (2007). Skype financials: a few more notes on the data. *Skype Journal*, No. 4. Prieiga per internetą: http://www.skypejournal.com/blog/2007/04/ [žiūrėta 2007-06-26].
- GRIGAS, G.; PEDZEVIČIENĖ, S. (2007). Mūsų asmenvardžiai internete. *Kompiuterija*, nr. 4, p. 27.
- ISO/IEC 9995 (1994). Information technology – keyboard layouts for text and office systems. Part 1–8. ISO.


## SPELLING ERRORS OF NAMES IN INSTANT MESSENGERS REGISTRATION DATA


**Gintautas Grigas, Sigita Pedzevičienė**



S u m m a r y

The name misspelling statistics in login data of instant messenger Skype is presented. Frequently occured errors of changing letters with diacritic marks by similar letters of English (ASCII) alphabet are investigated in Czech, Danish, Estonian, German, Icelandic, Latvian, Lithuanian, and Polish languages. Six most popular names from every language and three largest city names from every country and containing letters with diacritics are taken into account. It is found that Danishes and Germans write practically without errors, the most errors were made by Lithuanians. The hypotheses about the reasons of errors are presented.